A Workflow Demonstrator for Processing Catalysis Research Data

Abraham Nieva de la Hidalga[1,2†], Donato Decarolis[1,2], Shaojun Xu[1,2], Santhosh Matam[1,2], Willinton Yesid Hernández Enciso[1,2], Josephine Goodall[1,2], Brian Matthews[3] & C. Richard A. Catlow[1,2,4]

[1]UK Catalysis Hub, Research Complex at Harwell, Rutherford Appleton Laboratory, R92 Harwell, Oxfordshire OX11 0FA, UK

[2]School of Chemistry, Cardiff University, Main Building, Park Place, Cardiff, CF10 3AT, UK

[3]Scientific Computing Department STFC, Rutherford Appleton Laboratory, Harwell Campus, Didcot, OX11 0QX, UK

[4]Department of Chemistry, University College London, London WC1E 6BT, UK

Keywords: Workflow demonstrator; Prototyping; Scientific workflow Management; Catalysis research data; High-throughput processing and analysis

Citation: Nieva de la Hidalga, A, et al.: A workflow demonstrator for processing catalysis research data. Data Intelligence 4(2), 455-470 (2022). doi: 10.1162/dint_a_00143

Received: August 17, 2021; Revised: November 11, 2021; Accepted: February 5, 2022

ABSTRACT

The UK Catalysis Hub (UKCH) is designing a virtual research environment to support data processing and analysis, the Catalysis Research Workbench (CRW). The development of this platform requires identifying the processing and analysis needs of the UKCH members and mapping them to potential solutions. This paper presents a proposal for a demonstrator to analyse the use of scientific workflows for large scale data processing. The demonstrator provides a concrete target to promote further discussion of the processing and analysis needs of the UKCH community. In this paper, we will discuss the main requirements for data processing elicited and the proposed adaptations that will be incorporated in the design of the CRW and how to integrate the proposed solutions with existing practices of the UKCH. The demonstrator has been used in discussion with researchers and in presentations to the UKCH community, generating increased interest and motivating further development.

† Corresponding author: Abraham Nieva de la Hidalga (Email: nievadelahidalgaa@cardiff.ac.uk; ORCID: 0000-0001-7348-7612).

1. INTRODUCTION

Experimental and computational simulation techniques developed to understand the nature of materials and their practical applications in catalysis research rely on the use of data for building and validating complex models (such as the example in Figure 1). The UK Catalysis Hub (UKCH) enables cutting-edge research in catalytic science, by facilitating access to state-of-the-art resources and expertise. UKCH provides access to well equiped laboratories, central facilities provided by the Science and Technology Facilities Council (STFC) and offers expert advice for processing and analysis of the data produced from experiments and theoretical models.

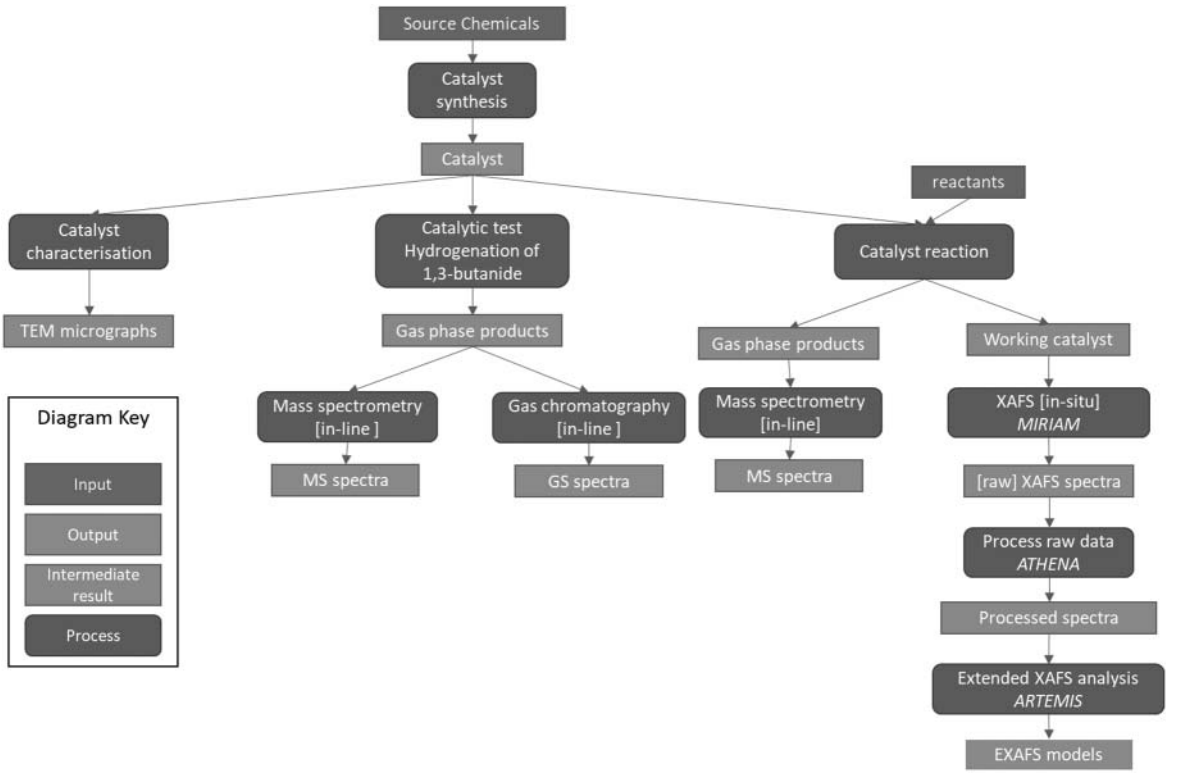


Figure 1. Diagram of an in-situ XAFS analysis experiment [11]. The target of this proposal are the processes and outputs after XAFS microspectroscopy[①], on the lower rightmost branch of the experimental process using ATHENA and ARTEMIS for processing and analyses of XAFS data.

UKCH researchers use advanced processing and analysis software such as Mantid [3], DAWN [4], Larch [29], and Demeter [33] to handle the data produced by their research projects. These tools allow scientists to process and analyze data interactively. Additionally, each scientist has a choice of analysis software such as MATLAB, R, and Excel, to further analyze data and to format results for publishing. STFC

① Performed at Diamond Beamline B22: Multimode InfraRed Imaging And Microspectroscopy (MIRIAM).

facilities (CLF [7, 8], Diamond [13, 20], and ISIS [15, 21]) operate 24 hours a day and have the capacity to performing thousands of readings which produce large datasets that require further processing and analysis. Naturally, the time employed in processing and analyzing data increases with the size of the datasets. Moreover, new experiment proposals aiming to collect even larger quantities of data push the boundaries of the processing capacity of analysis tools [37].

Having in mind the current and future requirements for processing and analysis of increasing data volumes, the UKCH started designing a virtual research environment, the Catalysis Research Workbench (CRW). The development of this platform requires identifying the processing and analysis needs of the UKCH members and mapping them to potential solutions. In this requirements collection phase, the UKCH implemented a workflow demonstrator to foster further discussion and analysis of the requirements for the CRW. The goal of the demonstrator is to introduce the concept of managed scientific workflows and discuss their integration in the day-to-day practices of UKCH researchers. The demonstrator has been used in discussion with researchers and in presentations to the UKCH community, generating increased interest and motivating further development.

2. RELATED WORK

The use of software prototypes is an established software engineering practice [6, 22, 28, 34, 35, 36]. A demonstrator is a type of functional prototype which is used in proof-of-concept studies to support the illustration of complex design proposals to a wide range of system stakeholders. The demonstrator can be presented by the designer who describes the details of the implementation while performing a specific set of tasks, often scripted, and then requests feedback from the user community.

There are various cases in which prototypes (and demonstrators) have been used successfully to present implementation proposals and to refine and prioritize user requirements. Prototyping has been used for multiple purposes such as the description of architechtural decisions, discussion of interface design, and presentation of new functionalities. Davis et al. use a Web service-based e-science demonstrator to explain the architectural design for a text mining platform [10]. Klampanos et al. describe the implementation of an information registry prototype to demonstrate how it can enable collaboration and ensure consistency across the distributed infrastructure for Dispel and dispel4py [23]. Leong et al. present the implementation of three use cases to demonstrate the feasibility and benefits of applying a cloud driven approach to supercomputing ecosystems [25] for large scale experimental facilities.

In the workflow domain, Goble et al. used a demonstrator to present the design principles and functionality of the myGrid middleware suite, to facilitate the work of bioinformaticians [17]. Nieva et al. describe the use of different prototypes in the review of alternative designs for a web interface for the Taverna workflow management system [28]. Watkins et al. present Workspace, a scientific workflow system that includes rapid prototyping features enabling the testing of different components and configurations during the design of complex workflows [38].

3. PROBLEM FORMULATION

Large scale research facilities such as the Central Laser Facility (CLF [7, 8]) Diamond Light Source (Diamond [13, 20]), and ISIS Muon and Neutron Source (ISIS [15, 21]) have an operational framework supported by their Data Management Policies. This framework governs their Laboratory Information Management (LIM) systems and the Data Management System (DMS). The main commonality of these facilities is that they use ICAT, an advanced catalogue system that combines LIM and DMS functionalities [16]. ICAT is developed by the Scientific Computing Department of the Science and Technology Facilities Council (SCD-STFC) and other institutions. The ICAT system contains complementary data for each experiment like proposal, PI, Experimenter, Grant(s), device(s), experiment metadata and experiment results. As a result, the extended workflows of CLF, Diamond and ISIS can be generalized as shown in Figure 2.

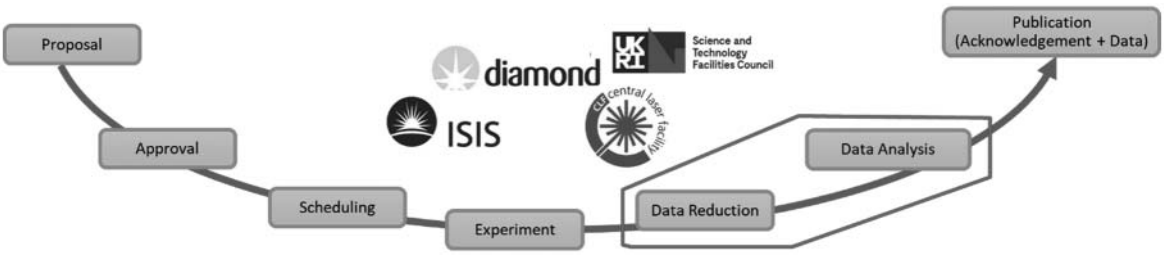


Figure 2. The generic processing workflow of CLF, Diamond, and ISIS (Adapted from [27]). The data reduction and analyses tasks highlighted in red are traditionally performed by facilities users, decoupled from facilities.

As Figure 2 indicates, Data Reduction and Data Analyses tasks are entrusted to the facilities users, i.e., scientists who have been awarded experimental time at the facilities. These tasks are the ones which require further support, as researchers report that processing and analysing data after the experiment requires substantial amount of time and processing resources. The research facilities provide software for collecting and formatting the data generated (for instance Mantid [3] and DAWN [4]), however, the researchers still need to handle the data and combine it with other data according to their objectives. Researchers rely on a combination of data and software resources (own and shared) in their daily work. In this context, there are several issues that the researcher needs to handle, such as mastering the use of several types of analysis tools including lab equipment, processing software and databases; converting data so that it can be used at different stages; and ensure the reproducibility of the results by tracking equipment and software used, entry parameters, intermediate results, and versions of completed runs.

4. THE PROPOSED APPROACH

The need for supporting users in the processing and analysis of research data has gained higher priority because the size and complexity of datasets is constantly increasing. This is the case of XAS analysis with the development of higher throughput analysis devices [37] and longer running times. Up to now, researchers have managed using interactive software for formatting, processing, reducing, and summarizing experimental data. However, researchers are spending longer hours processing and formatting data, which distracts them

from their experimental work. The target of the workflows proposed are these time-consuming activities. We aim to build on the experience gathered in the adoption of workflow technologies and proposed the creation of concrete examples which demonstrate the advantages of using scientific workflow management tools when compared to current processing practices.

4.1 Example for the Workflow Demonstrator

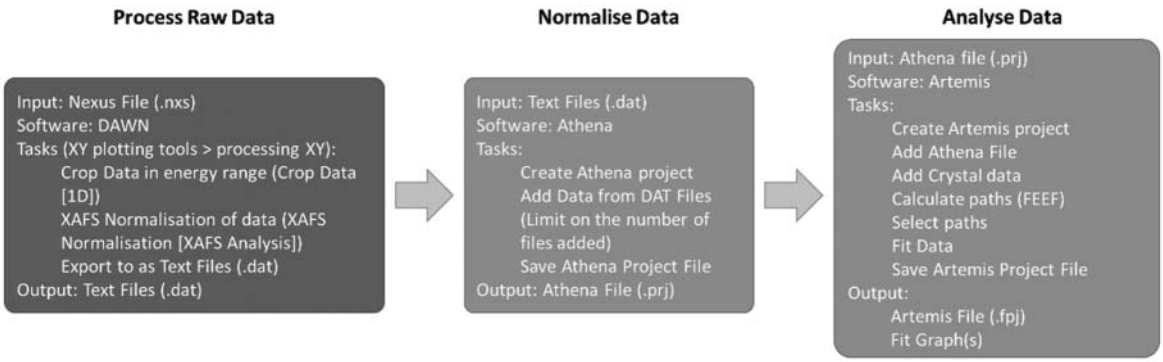
The explicit definition of processing workflows provides a complete view of the activities performed, the software used, and the data consumed and produced. After defining the workflow, its individual tasks can then be implemented modularly, allowing the combination, and swapping of components. The processing X-ray Absorption Spectroscopy (XAS) data is relevant because of the number of experiments performed and the quantities of data produced. Normally a scientist use Artemis and Athena [33] in a well-defined structured fashion. Moreover, the XAS processing workflow is well documented and there are several examples and tutorials on the use of Artemis and Athena for performing the workflow tasks [31, 33]. Athena and Artemis tasks can be scripted in Perl using Demeter. Additionally, there are alternative tools which have been proposed and can also be automated through scripting (e.g., Larch [29]). All these considerations made the XAS processing workflow the selected target to implement as an example for the demonstrator.

4.2 The XAS Processing Workflow

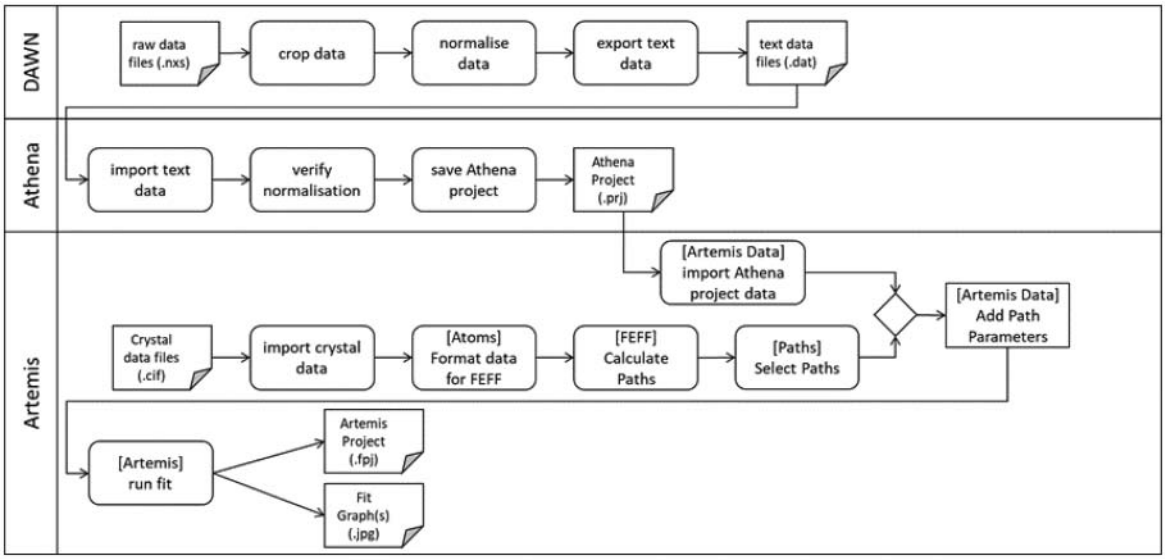
The XAS processing workflow consists of three tasks: Process Raw Data, Normalise Data, and Analyse Data. This division of the tasks is derived from the Ravel's online courses [31, 32], the DAWN tutorials [14] and from discussion with coauthors about processing practices. Figure 3A presents an overview of the three tasks of the XAS processing workflow. At this level, we can name the software, inputs, and outputs for each task of the workflow. The analysis of the workflow can be further refined to identify the sub-tasks within each task, providing a modular view of the workflow components. Figure 3B shows a finer grained description of the sub-tasks of the workflow. At this level, tasks are better defined as modules which can be implemented independently. This representation of the subtasks, including their relationships and precedence, including the inputs, resources, and outputs is the stating point for the implementation of the workflow.

4.3 Implementing the Workflow

After identifying the core tasks to implement, sequence, as well as the expected inputs and outputs from each sub-task, it was possible to decide the alternative ways to implement the workflow and to define which metrics to use to analyse the performance of the different instances. Three types of workflow configurations have been implemented: Manual (Interactive/User driven), Scripted (automated with scripts), and Managed (Semi-automated/User supervised). The manual workflow is just a reproduction of the textbook example, using the sample data from literature [31, 32] and repeated using an example from our experimental colleagues. The manual version is also used to calculate the baseline time for execution of one full cicle of the workflow, from raw data to fitted data results. The two versions of the scripted workflow were



A. Overview



B. Detailed view

Figure 3. Overview and detailed view of the XAS processing workflow.

implemented using Demeter and Larch (one for each). The Demeter version is scripted in Perl and allows running the same process as the manual workflow. The main difference is that the interface is text based and the operations are presented in a text menu. The Larch version of the scripted workflow is implemented using Larch and Jupyter Notebooks (Python). Finally, two managed versions of the workflow were designed to be executed using Nextflow [12] in combination with Larch and Demeter.

The first three versions of the workflow were fully implemented and used in demonstrations while the Nextflow managed version is in the process of being implemented for execution on a high-performance computing environment.

5. EXPERIMENTS AND ANALYSIS

Each of the three alternative implementations of the workflow was designed and tested using sample data from the textbook examples. The comparison of processing times was then made using two datasets containing data form actual UKCH experiments. The running times were then averaged and used for comparing the proposals and demonstrated to scientists to get feedback on the implementation.

5.1 Datasets and Experimental Setup

As described previously, three datasets were used. The first dataset is used for development and testing and is the dataset of Ravel's textbook example [31, 32]. The textbook dataset consists of one file containing the crystallographic data for Iron Sulfide (pyrite, FeS2) and a transmission scan of FeS2 taken at room temperature at beamline 13BM at the Advanced Photon Source [32]. The other two example datasets consist of a nexus file from containg Rh4CO spectral data gathered from the I20 Energy Dispersive EXAFS (EDE) beamline at Diamond Lightsource and a crystallographic data file of tetrarhodium dodecacarbonyl obtained from the Crystallographic Open Database [9, 18].

The software used for implementing the workflows included DAWN V.2.16.1, Demeter V. 0.9.26 (which includes Artemis and Athena), Larch V. 0.9.47, Perl V. 5.12.3, Python V. 3.6.10, and Jupyter V. 6.0.3. The system used for running the experiments was a laptop computer with Windows 10 64-bit operating system, Intel Core i5-8250 1.60 GHz Processor, and 8 GB Memory.

5.2 Overall Comparison of Results

The three implemented examples of the workflow were individually timed for comparison of potential for speeding up the processing and analysis of XAS data. The manual version of the workflow based on the textbook example takes about 24 minutes to produce one complete run from raw data to fitted data results. This average time was taken from performing the workflow activities manually with ten samples of the Rh4CO spectra and then averaging the processing time from start to finish. Using these data, we calculate that processing a dataset of 3,790 readings would take about 63 days. The experts in the group consider that they can perform one complete run in 10 minutes, which would require approximately 23 days to process the 3,790 readings dataset.

The first scripted version of the workflow uses Demeter and Perl and it allows fast processing in about 22 hours for a dataset of 3,790 groups (~1 day). This is a considerable improvement from the manual workflow. The second scripted version of the workflow uses Larch, Jupyter and Python. It is slower than the Demeter version, but still can reduce the processing time to 103 hours (4.3 Days), taking only 20% of the time required for manually processing a full dataset.

The results in Table 1 were obtained using a laptop computer with limited memory and processor. In comparison, the initial results of a NextFlow-Larch version of the workflow reduced processing time to 7 hours and 21 minutes for the largest dataset (4000 groups) when executed in the ARCCA-HPC cluster. At this stage, the presentation of the demonstrator to stakeholders indicates that the approach could be applied to real life scenarios, as positive reviews and suggestions for improvement indicate.

Table 1. Comparison of worflow instances in terms of speed.

Manual Workflow	Task	Software	Time	Input	Output
	Process raw data	DAWN	8 min	1 nexus [.nxs] file	3580 – 4000 files
	Normalise data	Athena	3 min	1 data [.dat] file	1 Athena file
	Analyse data	Artemis	21 min	1 Athena [.prj] file 1 Crystal [.inp/.cif] file	1 Artemis file
	Novice user processing 1 dataset		**~63 days**	**~ 24 mins to produce 1 fit**	
	Expert processing 1 dataset		~26 days	~ 10 mins to produce 1 fit	
Demeter-Perl Scripted Workflow	Task	Software	Time	Input	Output
	Process raw data	DAWN	8 min	1 nexus [.nxs] file	3580 – 4000 files
	Normalise data	Demeter	64 min	500 data [.dat] files	500 Athena files
	Analyse data	Demeter	21 min	500 Athena [.prj] file 1 Crystal [.inp/.cif] file	500 Demeter [.dpj] files 500 Fit [.fit] files 500 Log files
	Processing a dataset with 3,790 groups		**~22 Hours**	**~ 21 sec. to produce 1 fit**	
Larch-Jupyter Scripted Workflow	Task	Software	Time	Input	Output
	Process raw data	DAWN	8 min	1 nexus [.nxs] file	3580 – 4000 files
	Normalise data	Larch	8 min	4000 data [.dat] files	4000 Athena files
	Analyse data	Larch	814 min	500 Athena [.prj] file 1 Crystal [.inp/.cif] file	500 Demeter [.dpj] files 500 Fit [.fit] files 500 Log files
	Processing a dataset with 3,790 groups		**~103 Hours**	**~ 1.5 min. to produce 1 fit**	

5.3 Results Analysis

The three versions of the workflow have been showcased and discussed with researchers in two separate occasions, providing valuable feedback, suggestions for improvement and future developments. The workflows are not intended to be fully operational processing and analysis tools, instead the functionalities and details of the examples is intended to illustrate the benefits of adopting a workflow-oriented design approach. The workflows were first demonstrated at a workshop with our coauthors and served to demonstrate the feasibility of automating repetitive tasks and provided some recommendations for improvements for the workflows. The second presentation of the demonstrator during one of the monthly UKCH seminars, exposed the workflows to a larger community and prompted for suggestions and queries about implementing other analyses using workflows.

At this stage we can highlight the advantages and disadvantages of each of the workflow implementations, including the ones which are still under development. The scripted and managed versions of the workflows are faster for the processing and analysis of data. Moreover, expert users recommended improvements such as monitoring output values to determine if the executions should be terminated early.

Table 2. Comparison of workflow instances.

Workflow	Software	Type	Issues	Advantages
Manual	Artemis, Athena	Manual	Slow processing Limit on number of datasets loaded Individually processing datasets	Interactive visual interface Fine tuning control
Demeter	Demeter, Perl	Scripted	Limited by processing resources Text interface	Processing large quantities of data
Larch	Larch, Python, Jupyter Notebook	Scripted	Limited by processing resources Requires Demeter for one key task	Interactive visual interface Fine tuning control Processing large quantities of data
Nextflow 01	Demeter, Perl, Nextflow	Managed	Limited by processing resources Text interface	Processing large quantities of data Unsupervised execution
Nextflow 02	Larch, Python, Nextflow	Managed	Limited by processing resources Text interface Requires Demeter for one key task	Processing large quantities of data Unsupervised execution

5.3 Relevance of the Canonical Workflow Framework for Research

The Canonical Workflow Framework for Research (CWFR) is aimed at facilitating the interoperation of data management workflows across institutional boundaries [19]. In order to achieve this, the CWRF aims to explicitly document the repetitive tasks which are common across diverse institutional data management workflows (Figure 4). For the management and exploitation of Catalysis Research data, the CWRF model allows stepping back and looking at possibilities for integrating the facilities workflows to the workflows of other institutions accessing the facilities. In Figure 5, the diagram shows how the STFC workflow is aligned with the workflows of other institutions, and how the tasks of these workflows can be mapped to the CWRF tasks.

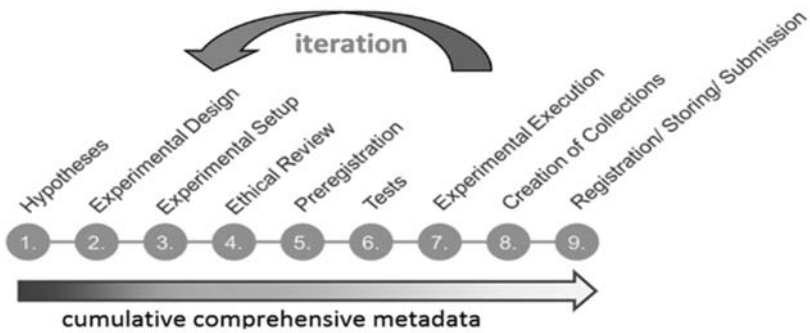


Figure 4. Common tasks identified in the first version of the CWFR (Adapted from [19]). The tasks are not all carried by one institution, they are complementary and can be fulfilled by different institutions collaborating in the research effort. This is shown in the example provided in Figure 5.

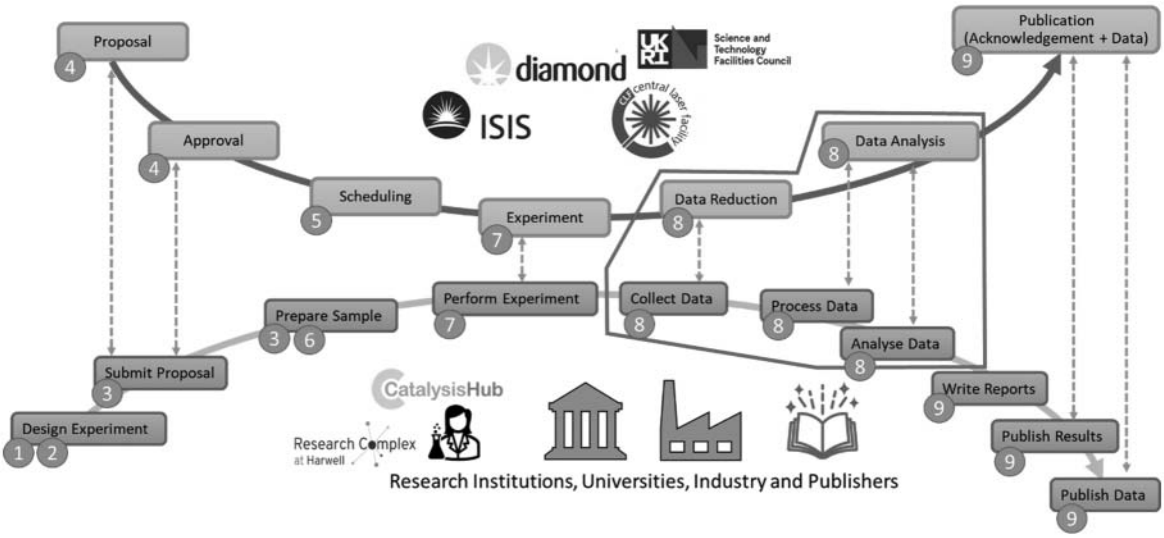


Figure 5. Parallels between the experimental workflows of STFC facilities and other institutions mapped to CWFR Tasks. The fiure shows two workflows, and the activities performed in parallel during research collaborations. The upper workflow is the same presented in Figure 2, while the lower one stands for the workflow performed by institutions accessing and collaborating with STFC facilities. The numbers in orange represent the tasks identified in the CWFR (see Figure 4). The five tasks highlighted in red correspond to activities supported by the type of workflows described in this paper.

The extended workflow for STFC facilities provides the basic scaffolding for integrating with other workflows. The main tool underpinning this workflow is the ICAT system [16]. ICAT combines the functionalities of Laboratory Information Management (LIM) system and the Data Management System (DMS). ICAT registers complementary data for each experiment like proposal, PI, Experimenter(s), Grant(s), device(s), experiment metadata and experiment results. ICAT is common to many facilites (UK and Overseas). ICAT supports the required management functionalities implementing the Core Scientific MetaData model (CSMD) [26]. This model captures metadata about the experiments and datasets produced at the facilities managed by STFC. By design, the operational workflows of the facilities rely on the CSMD, for managing experiments from the proposal stage to the collection and distribution of experimental data.

Additionally, the CSMD can enable the alignment of workflows of other institutions by mapping to other ontologies, such as PROV-O for tracking provenance [24], DCAT for the description of data objects [2], and SPAR [30] and SCHOLIX [5] for linking data objects and publications.

6. CONCLUSION AND FUTURE WORK

The implementation of the demonstrator with three versions of the XAS workflow, is a first attempt to promote greater usage of the Scientific Workflow approach at UKCH. This first version of demonstrator has stimulated the interest for further research on workflow management platforms. We plan to continue the

development by completing the nextflow version [12] and possibly adding examples for Galaxy [1], and Taverna [39], which provide different benefits. These will be then evaluated in further demonstrations to gather more requirements for implementation.

Looking forward, the UKCH will try to standardize the procedures for describing and implementing other processing workflows to support data processing and analysis. For this, we are considering new examples, such a Quasi-Elastic Neutron Scattering (QENS) and X-Ray Powder Diffraction (XRD) processing workflow. In the longer term, the evaluation of workflow implementation alternatives will help the UKCH in better defining the requirements and design constraints to be followed for the development of the Catalysis Research Workbench (CRW).

ACKNOWLEDGEMENTS

UK Catalysis Hub is kindly thanked for resources and support provided via our membership of the UK Catalysis Hub Consortium and funded by EPSRC grant: EP/R026939/1, EP/R026815/1, EP/R026645/1, EP/R027129/1 or EP/M013219/1(biocatalysis)). We acknowledge the support of provided by Advanced Research Computing at Cardiff (ARCCA) for the implementation and testing the NextFlow version of the workflow, ARCCA is part of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government.

AUTHOR CONTRIBUTIONS

A. Nieva de la Hidalga (nievadelahidalgaa@cardiff.ac.uk) contributed to the implementation and development of the workflow versions (Published at https://github.com/UK-Catalysis-Hub/XAS-Workflow-Demo). D. Decarolis (decarolisd@cardiff.ac.uk) contributed to the definition of the problem and provided the experimental data used to evaluate the models. S. Xu (XuS25@cardiff.ac.uk), S. Matam (MatamS@cardiff.ac.uk), W.Y. Hernández Enciso (yesidhdz@hotmail.com), and J. Goodall (josie.goodall@rc-harwell.ac.uk), collaborated in discussions about the design and functionalties provided. B. Matthews (brian.matthews@stfc.ac.uk) and C.R.A. Catlow (c.r.a.catlow@ucl.ac.uk) provided feedback on the design of the experiments. All authors reviewed the paper and provided further ideas for improving its contents.

REFERENCES

[1] Afgan, E., et al.: The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46 (W1), W537–W544 (2018)

[2] Albertoni, R., et al.: (2020) Data Catalog Vocabulary (DCAT)—Version 2. W3C Recommendation 4 February 2020. Available at: https://www.w3.org/TR/vocab-dcat-2/. Accessed 15 August 2020

[3] Arnold, O., et al.: Mantid—Data analysis and visualization package for neutron scattering and μ SR experiments. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 764, 156–166 (2014)

[4] Basham, M., et al.: Data analysis workbeNch (DAWN). Journal of Synchrotron Radiation 22, 853–858 (2015)

[5] Burton, A., et al.: The Scholix framework for interoperability in data-literature information exchange. D-Lib Magazine 23, Number 1/2 (2017)

[6] Clements, P.C.: Active reviews for intermediate designs (CMU/SEI-2000-TN-009). Available at: http://www.sei.cmu.edu/library/abstracts/reports/00tn009.cfm. Accessed 15 August 2020

[7] Central Laser Facility (CLF). Available at: https://www.clf.stfc.ac.uk/Pages/home.aspx. Accessed 15 August 2020

[8] Cooke, E.: The Central Laser Facility at 40. Available at: https://www.clf.stfc.ac.uk/Pages/The-Central-Laser-Facility-at-40.aspx. Accessed 15 August 2020

[9] Crystallographic Open Database: Structural analyses of tetracobalt dodecacarbonyl and tetrarhodium dodecacarbonyl, crystallographic treatments of a disordered structure and a twinned composite. Available at: https://www.crystallography.net/cod/cif/4/34/45/4344516.cif. Accessed 3 March 2020

[10] Davis, N., et al.: Web service architectures for text mining: An exploration of the issues via an e-science demonstrator. International Journal of Web Services Research (IJWSR), 3(4), 95–112 (2006)

[11] Decarolis, D., et al.: Effect of particle size and support type on Pd Catalysts for 1, 3-Butadiene Hydrogenation. Topics in Catalysis 61(3–4), 162–174 (2018)

[12] Di Tommaso, P., et al.: Nextflow enables reproducible computational workflows. Nature Biotechnology 35(4), 316–319 (2017)

[13] Diamond light source. Available at: https://www.diamond.ac.uk/Home.html. Accessed 15 August 2020

[14] Filik, J., Snow, T.: (2021) An introduction to DAWN—Tutorials. Available at: https://diamondlightsource.atlassian.net/wiki/spaces/DT/pages/1378501/Tutorials. Accessed 15 August 2020

[15] Findlay, D.J.S.: ISIS-pulsed neutron and muon source. In: 2007 IEEE Particle Accelerator Conference (PAC), pp. 695–699 (2007)

[16] Flannery, D., et al.: ICAT: Integrating data infrastructure for facilities-based science. In: 2009 Fifth IEEE International Conference on e-Science, pp. 201–207 (2009)

[17] Goble, C., et al.: The myGrid project: Services, architecture and demonstrator. In: Proceedings of the UK e-Science Programme All Hands Conference, pp. 595–603 (2003)

[18] Gražulis, S., et al.: Crystallography Open Database (COD): An open-access collection of crystal structures and platform for world-wide collaboration. Nucleic Acids Research 40, D420–D427 (2012)

[19] Hardisty, A.R., Wittenburg, P.: (2020) Canonical workflow framework for research CWFR. Position paper. Available at: https://codata.org/wp-content/uploads/2021/01/CWFR-position-paper-v3.pdf. Accessed 4 January 2021

[20] Helmers, C., Overman, H.G.: My precious! The location and diffusion of scientific research: evidence from the synchrotron diamond light source. The Economic Journal 127(604), 2006–2040 (2017)

[21] ISIS: ISIS neutron and muon source. Available at: https://www.isis.stfc.ac.uk/Pages/home.aspx. Accessed 15 August 2020

[22] Käpyaho, M., Kauppinen, M.: Agile requirements engineering with prototyping: A case study. In: 2015 IEEE 23rd International Requirements Engineering Conference, pp. 334–343 (2015)

[23] Klampanos, I.A., Martin, P., Atkinson, M.: Consistency and collaboration for fine-grained scientific workflow development: The dispel4py information registry. Technical Report (2019)

[24] Lebo, T., et al.: (2013). PROV-O: The PROV Ontology, W3C Recommendation 30 April 2013. Available at: https://www.w3.org/TR/prov-o/. Accessed 25 October 2020

[25] Leong, S.H., et al.: SELVEDAS: A data and compute as a service workflow demonstrator targeting supercomputing ecosystems. In: 2020 IEEE/ACM International Workshop on Interoperability of Supercomputing and Cloud Technologies (SuperCompCloud), pp. 7–13 (2020)

[26] Matthews, B., Fisher, S.: CSMD: The core scientific metadata model, CSMD 4.0 reference. document. Available at: http://icatproject-contrib.github.io/CSMD/csmd-4.0.html. Accessed 20 September 2019

[27] Matthews, B.: The FAIR experiment facilities science. Presentation during the 17th RDA Plennary, FAIR Publishing of Chemistry Research Data Objects. Presented on April 20, 2021.
[28] Nieva de la Hidalga, A., Hardisty, A., Jones, A.: SCRAM–CK: Applying a collaborative requirements engineering process for designing a Web based e-science toolkit. Requirements Engineering 21(1), 107–129 (2016)
[29] Newville, M.: Larch: An analysis package for XAFS and related spectroscopies. Journal of Physics: Conference Series 430, Article No. 012007 (2013)
[30] Peroni, S., Shotton, D.: The SPAR ontologies. In: International Semantic Web Conference, pp. 119–136 (2018)
[31] Ravel, B.: Bruce Ravel XAS course 2011. Available at: https://www.diamond.ac.uk/Instruments/Spectroscopy/Techniques/XAS.html. Accessed 15 August 2020
[32] Ravel, B.: (2013) Bruce Ravel's XAS education materials, step by step tutorial. Available at: https://github.com/bruceravel/XAS-Education/tree/master/Examples/FeS2. Accessed 15 August 2020
[33] Ravel, B., Newville, M.: ATHENA, ARTEMIS, HEPHAESTUS: Data analysis for X-ray absorption spectroscopy using IFEFFIT. Journal of Synchrotron Radiation 12, 537–541 (2005)
[34] Sutcliffe AG: User-centered requirements engineering. Springer, London (2002)
[35] Sutcliffe AG: (2009) On the inevitable intertwining of requirements and architecture. In: Lyytinen, K., et al. (eds.) Design Requirements Engineering: A Ten-Year Perspective, pp.168–185. Springer, Berlin (2009)
[36] Teixeira, L., et al.: Requirements engineering using mockups and prototyping tools: Developing a healthcare Web-application. In: International Conference on Human Interface and the Management of Information, pp. 652–663 (2014)
[37] Xu, S.J., et al.: High throughput XAS reactor system for operando spectroscopy study. Presentation at UK Catalysis Conference 2021 on January 8, 2021
[38] Watkins, D., et al.: Workspace-A scientific workflow system for enabling research impact. In: MODSIM2017: The 22nd International Congress on Modelling and Simulation, (2017)
[39] Wolstencroft, K., et al.: The Taverna workflow suite: Designing and executing workflows of Web Services on the desktop, Web or in the cloud. Nucleic Acids Research 41(W1), W557–W561 (2013)

AUTHOR BIOGRAPHY

Abraham Nieva de la Hidalga (Cardiff University) is a Research Associate in data management and software development at UK Catalysis Hub. He received a Ph.D. degree from the School of Computer Science, University of Manchester, in 2010. He has collaborated in research projects on semantic filtering of commercial emails (Commius 2010) workflow patterns for end-user-built Web-service mashups (SOA4ALL 2011); interfaces for execution and modification of workflows (BioVeL 2012); designing a biotechnology virtual laboratory (UPP 2015); modelling environmental research infrastructures (ENVRI 2016); and quality management standards for specimen's digitisation workflows and rapid 3D digitization (2018).
ORCID: 0000-0001-7348-7612

Donato Decarolis obtained a Chemistry Ph.D. degree from University College London, researching the effect of precious metal nanoparticle size and support type on catalytic activity using in-situ spectroscopy methods, particularly X-ray absorption spectroscopy (XAS). He joined the Centre for Nanoporous Materials to study the absorption of CO and N2 of Metal-organic frameworks using XAS methods (BP collaboration 2017). He worked in the Synchrotron Techniques for African Research and Technology (START) project (University of Southampton 2018). He joined UK Catalysis Hub (2019) to research the structure-activity relationship of catalysts using operando characterisation techniques.

Shaojun Xu obtained his Chemistry Ph.D. degree from the University of Manchester, researching mechanisms of plasma-catalysis for C1 compounds applying in-situ spectroscopic methods; his work was recognized with the International Plasma Chemistry Society scholarship. After his Ph.D., he worked as postdoctoral researcher at the University of Manchester investigating the potential of pore structures and porosity of MOFs in heterogeneous catalysis. He joined Cardiff University and the UK Catalysis Hub (2019) to work on the development of in-situ and operando methods for heterogeneous and homogeneous catalysis. He is interested in the mechanics of chemical reactions for energy, innovation of reaction processes, and development of operando characterisation methods.

Santhosh Kumar Matam is a Research Associate at Cardiff University based at the UK Catalysis Hub. He earned his PhD from Humboldt University of Berlin, Germany. The Research Council of Norway and Swiss National Foundation grants helped him to pursue his research, which is centred on in-situ and operando spectroscopy for deriving catalyst structure-activity relationships. He employs Neutron, X-ray and Laser based techniques, collaborating with computational chemists to design and develop catalytic materials for energy and environmental applications. He is also interested in operando reactors that allow analysis of chemical processes without intrinsic limitations.

Willinton Yesid Hernández Enciso received his Ph.D. in Chemistry from the University of Seville (Spain) in 2010, working on the design and characterization of heterogeneous catalysts for the Preferential Oxidation of CO in presence of hydrogen (PROX reaction). Since obtaining his Ph.D. he has had opportunity to explore different research topics including diesel-exhaust-gas treatment, catalytic oxidation, hydrogen generation, and biomass valorisation at IRCELYON (in France), ICIQ (in Spain) and Ghent University (in Belgium) Solvay (China) and UK Catalysis Hub. Currently, he is a Postdoctoral Researcher at the ALBA Synchrotron on the MIRAS beamline.

Josephine Goodall is the Project Manager of the UK Catalysis Hub at Cardiff University and has interests in all of the aspects of catalysis within the Hub. She has a MEng in Materials Economics and Management from Oxford in 2004 and a Ph.D. in Chemistry from UCL in 2009. After completing Her Ph.D. she went on to investigate the synthesis of nano-ceramics as phosphor materials and the synthesis on bio-inspired nano-metal sulphide materials for the electro-catalytic reduction of CO2 at UCL.

Brian Matthews leads the Open Data Systems Group, in the Scientific Computing Department of the Science and Technology Facilities Council. He leads research and development projects concentrating on advance management, engineering, curation, and analysis of scientific data. Brian has 30 years of research and development experience in the management, preservation and analysis of scientific data and access to compute resources. He has worked in formal methods for software engineering, data modelling; web technology (W3C SKOS recommendation), grid and distributed systems, security, scientific data management, and data and software preservation.

Sir **C. Richard A. Catlow** (FRS FRSC FInstP) is a professor at University College London and Cardiff University, He is a funding member and PI of the UK Catalysis Hub. He has established experience in the development and applications of experimental and computer modelling techniques in catalysis and molecular sciences. He has extensive experience in the field of HPC simulation techniques. He has been PI of the EPSRC funded Materials Chemistry HPC consortium for 15 years and has wide experience in managing large flexible consortium grants including a portfolio partnership grant (2005–2010), High-Performance Computing Consortium (2008–2013), and the Centre for Catalytic Science (2011–2016).